Mathmatial Foundations of Reinforcement Learning

Markov decision processes: dynamic programming

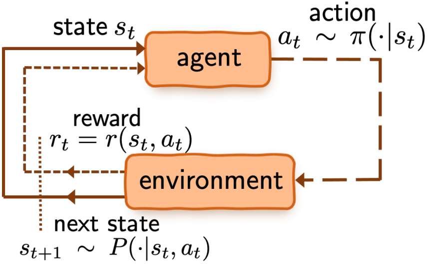
Yuting Wei

Statistics & Data Science, Wharton
University of Pennsylvania

Fall 2023

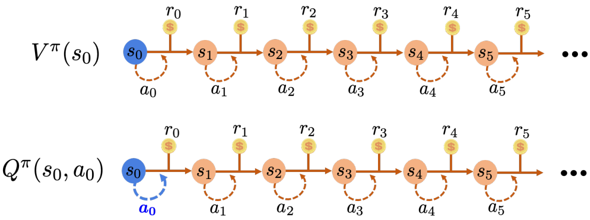
Outline

Infinite-horizon Markov decision process (MDP)



- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: immediate reward
- $\pi(\cdot|s)$: policy (or action selection rule), deterministic or random
- $P(\cdot|s, a)$: transition probabilities

Value function and Q-function



Value function of policy π: cumulative discounted reward

$$\forall s \in \mathcal{S}: \quad V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]$$

Q-function of policy π:

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad Q^\pi(s,a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a\right]$$
$$= \mathbb{E}\left[r(s,a)\right] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\pi(s')$$

Basic tasks

Policy evaluation:

- given a policy π, how good is it?

Basic tasks

Policy evaluation:

- given a policy π, how good is it?

Policy improvements:

- given a policy π, can we find a better one?

Basic tasks

Policy evaluation:

- given a policy π, how good is it?

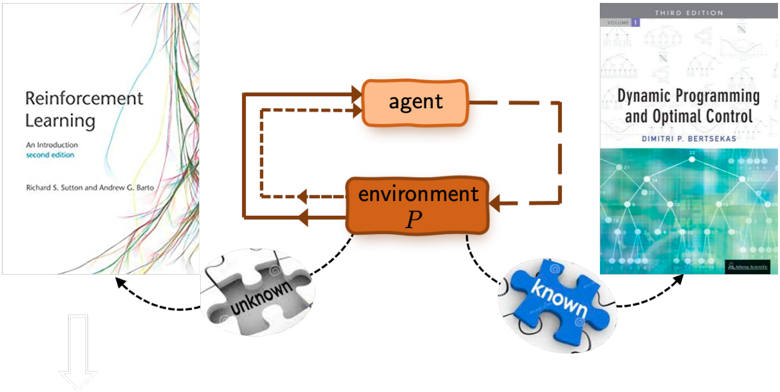
Policy improvements:

- given a policy π, can we find a better one?

Policy optimization:

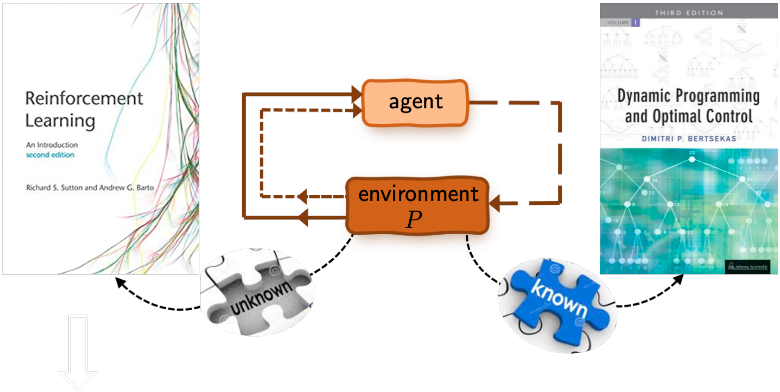
- can we find the best policy for the given MDP?

Planning versus learning



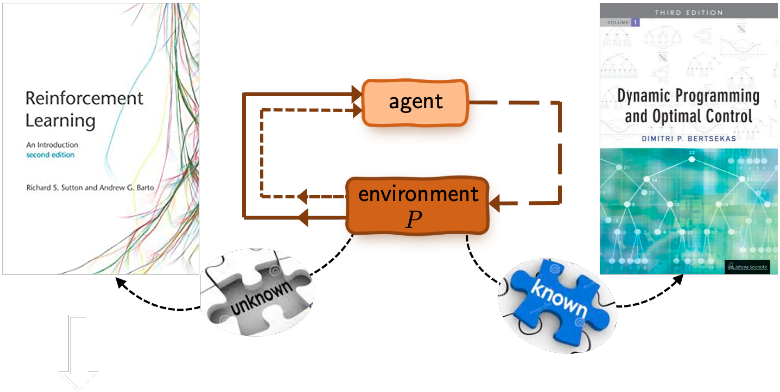
- **Planning:** solve for a desired policy given model specification

Planning versus learning



- **Planning:** solve for a desired policy given model specification
- **Learning:** learn a desired policy from samples w/o model specification

Planning versus learning



- **Planning:** solve for a desired policy given model specification
- **Learning:** learn a desired policy from samples w/o model specification

  We'll focus on planning first.

Policy improvement

Partial ordering of policies

Definition 1 (Partial ordering)

Define a partial order over policies: denote

$$\pi' \geq \pi \qquad \text{if} \quad \forall s \in \mathcal{S}, \quad V^{\pi'}(s) \geq V^{\pi}(s).$$

- The policy π' is an *improvement* over π since it improves its value in all states.

Partial ordering of policies

Definition 1 (Partial ordering)

Define a partial order over policies: denote

$$\pi' \geq \pi \qquad \text{if} \quad \forall s \in \mathcal{S}, \quad V^{\pi'}(s) \geq V^{\pi}(s).$$

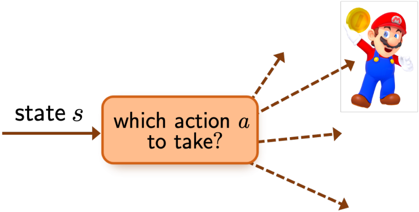
- The policy π' is an *improvement* over π since it improves its value in all states.

Question

Given a policy π, how to find an improved policy?

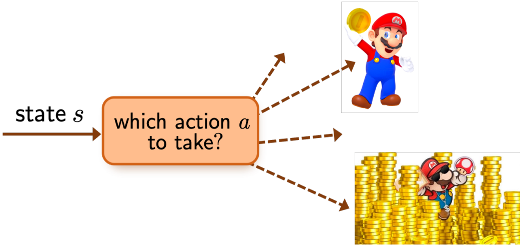
Policy improvement via one-step look-ahead

Given the Q-function Q^π of some policy π.



Policy improvement via one-step look-ahead

Given the Q-function Q^π of some policy π.



At each state s, can we identify an action a such that

$$Q^\pi(s,a) \geq V^\pi(s)?$$

- taking action a at state s leads to a higher cumulative reward than following policy π.

Policy improvement theorem

Theorem 2 (Policy improvement theorem)

Choose some stationary policy π, and let π' be a deterministic policy such that

$$\forall s \in \mathcal{S}, \qquad Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s).$$

Then $V^{\pi'} \geq V^{\pi}$, i.e., π' is an improvement over π.

Policy improvement theorem

Theorem 2 (Policy improvement theorem)

Choose some stationary policy π, and let π' be a deterministic policy such that

$$\forall s \in \mathcal{S}, \qquad Q^\pi(s, \pi'(s)) \geq V^\pi(s).$$

Then $V^{\pi'} \geq V^\pi$, i.e., π' is an improvement over π.

- Define the **greedy** policy w.r.t. some Q as

$$\pi_Q = \mathsf{Greedy}(Q), \quad \text{i.e.} \quad \pi_Q(s) = \arg\max_{a \in \mathcal{A}} Q(s, a).$$

Policy improvement theorem

Theorem 2 (Policy improvement theorem)

Choose some stationary policy π, and let π' be a deterministic policy such that

$$\forall s \in \mathcal{S}, \qquad Q^\pi(s, \pi'(s)) \geq V^\pi(s).$$

Then $V^{\pi'} \geq V^\pi$, i.e., π' is an improvement over π.

- Define the **greedy** policy w.r.t. some Q as

$$\pi_Q = \text{Greedy}(Q), \quad \text{i.e.} \quad \pi_Q(s) = \arg\max_{a \in \mathcal{A}} Q(s, a).$$

- The greedy policy $\pi' = \text{Greedy}(Q^\pi)$ w.r.t. Q^π is an improvement over π:

$$\begin{aligned} Q^\pi(s, \pi'(s)) &= \max_{a \in \mathcal{A}} Q^\pi(s, a) \\ &\geq \sum_{a \in \mathcal{A}} \pi(a|s) Q^\pi(s, a) = V^\pi(s) \qquad \Longrightarrow \qquad V^{\pi'} \geq V^\pi \end{aligned}$$

Proof of policy improvement theorem

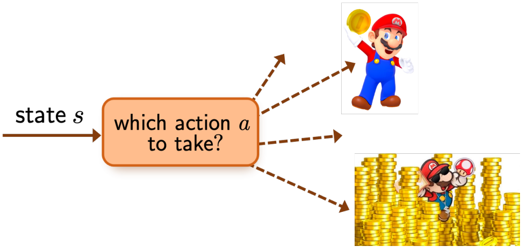
For each state $s \in \mathcal{S}$,

$$
\begin{aligned}
V^{\pi}(s) &\leq Q^{\pi}(s, \pi'(s)) \\
&= \mathbb{E}\left[r(s_0, \pi'(s_0)) + \gamma V^{\pi}(s_1) \,\middle|\, s_0 = s, a_0 = \pi'(s) \right] \\
&\leq \mathbb{E}\left[r(s_0, \pi'(s_0)) + \gamma Q^{\pi}(s_1, \pi'(s_1)) \,\middle|\, s_0 = s, a_0 = \pi'(s) \right] \\
&\leq \mathbb{E}\left[r(s_0, \pi'(s_0)) + \gamma r(s_1, \pi'(s_1)) + \gamma^2 V^{\pi}(s_2) \,\middle|\, s_0 = s, a_0 = \pi'(s) \right] \\
&\leq \cdots \\
&\leq \mathbb{E}\left[r(s_0, \pi'(s_0)) + \gamma r(s_1, \pi'(s_1)) + \gamma^2 r(s_2, \pi'(s_2)) + \cdots \,\middle|\, s_0 = s \right] \\
&\leq V^{\pi'}(s)
\end{aligned}
$$

One-step improvement leads to value increase.

Finding the optimal policy of MDPs

Optimal value and optimal policy

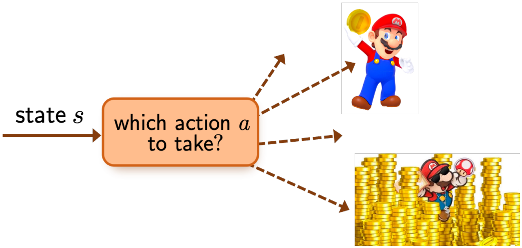


- **optimal value / Q function:**

$$V^\star(s) := \max_\pi V^\pi(s), \qquad Q^\star(s,a) := \max_\pi Q^\pi(s,a)$$

where the search is over all policies possibly non-stationary and random.

Optimal value and optimal policy



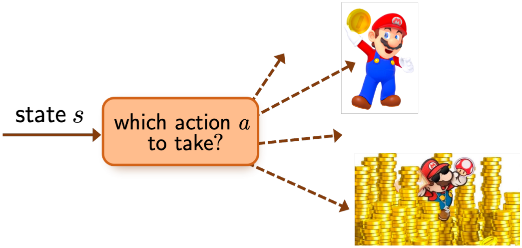
- **optimal value / Q function:**

$$V^\star(s) := \max_\pi V^\pi(s), \qquad Q^\star(s,a) := \max_\pi Q^\pi(s,a)$$

where the search is over all policies possibly non-stationary and random.

- **optimal policy** $\pi^\star$: the policy that maximizes the value function.

Optimal policy: existence



Lemma 3 ([Bellman, 1952])

For infinite-horizon discounted MDPs, there always exists a stationary and deterministic policy $\pi^\star$, such that for all $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$V^{\pi^\star}(s) = V^\star(s), \qquad Q^{\pi^\star}(s,a) = Q^\star(s,a).$$

- Using stationary and deterministic policies suffices.
- See [Agarwal et al., 2019] for a proof.

Bellman's optimality equations

Theorem 4 (Bellman's optimality equations)

The optimal value/Q functions are unique and related via

$$V^\star(s) = \max_{a \in \mathcal{A}} Q^\star(s,a),$$
$$Q^\star(s,a) = \mathbb{E}[r(s,a)] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s').$$

Furthermore, $\pi^\star = \pi_{Q^\star} = \mathrm{Greedy}(Q^\star)$ *is an optimal policy (tie-breaking arbitrarily).*

- Knowing the optimal Q-function allows us to find the optimal policy.
- The optimal values are unique, but the optimal policy is not necessarily unique.

Proof of Bellman's optimality equations

Proof of $V^\star(s) = \max_{a \in \mathcal{A}} Q^\star(s,a)$:

$$
\begin{aligned}
V^\star(s) &= \max_\pi V^\pi(s) \\
&= \max_\pi \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s,a)] \\
&\leq \max_\pi \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\star(s,a)] \\
&= \max_{a \in \mathcal{A}} Q^\star(s,a).
\end{aligned}
$$

Proof of Bellman's optimality equations

Proof of $V^\star(s) = \max_{a\in\mathcal{A}} Q^\star(s,a)$:

$$\begin{aligned} V^\star(s) &= \max_\pi V^\pi(s) \\ &= \max_\pi \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)] \\ &\le \max_\pi \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\star(s,a)] \\ &= \max_{a\in\mathcal{A}} Q^\star(s,a). \end{aligned}$$

On the other end, for any $a \in \mathcal{A}$, consider the (possibly non-stationary) policy that first takes action a and then follows $\pi^\star$. It follows that

$$V^\star(s) \ge Q^{\pi^\star}(s,a) = Q^\star(s,a).$$

This implies, by the arbitrariness of a,

$$V^\star(s) \ge \max_{a\in\mathcal{A}} Q^\star(s,a).$$

Proof of Bellman's optimality equations

Proof of $Q^\star(s,a) = \mathbb{E}[r(s,a)] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s')$**:**

$$\begin{aligned}
Q^\star(s,a) &= \max_\pi Q^\pi(s,a) \\
&= \max_\pi \Big[\mathbb{E}[r(s,a)] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\pi(s') \Big] \\
&= \mathbb{E}[r(s,a)] + \gamma \max_\pi \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\pi(s') \\
&= \mathbb{E}[r(s,a)] + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s')
\end{aligned}$$

Bellman's optimality principle

Bellman operator

$$\mathcal{T}(Q)(s,a) := \underbrace{\mathbb{E}[r(s,a)]}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \underbrace{\left[\max_{a' \in \mathcal{A}} Q(s',a') \right]}_{\text{next state's value}}$$

- one-step look-ahead

Bellman's optimality principle

Bellman operator

$$\mathcal{T}(Q)(s,a) := \underbrace{\mathbb{E}[r(s,a)]}_{\text{immediate reward}} + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \Big[\underbrace{\max_{a' \in \mathcal{A}} Q(s',a')}_{\text{next state's value}} \Big]$$

- one-step look-ahead

Bellman's optimality equation: $Q^\star$ is the *unique* fixed point to

$$\mathcal{T}(Q^\star) = Q^\star$$

Uniqueness is immediately implied by the γ-contraction on the next slide (verify!).

Richard Bellman

Contraction of the Bellman's operator

Lemma 5 (γ-contraction of Bellman operator)

For any Q and Q', it holds

$$\|\mathcal{T}(Q) - \mathcal{T}(Q')\|_\infty \leq \gamma\|Q - Q'\|_\infty.$$

Contraction of the Bellman's operator

Lemma 5 (γ-contraction of Bellman operator)

For any Q and Q', it holds

$$\|\mathcal{T}(Q) - \mathcal{T}(Q')\|_\infty \le \gamma \|Q - Q'\|_\infty.$$

Proof: $\|\mathcal{T}(Q) - \mathcal{T}(Q')\|_\infty$

$$= \gamma \max_{s,a} \left| \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q(s',a') \right] - \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q'(s',a') \right] \right|$$

Contraction of the Bellman's operator

Lemma 5 (γ-contraction of Bellman operator)

For any Q and Q', it holds

$$\|\mathcal{T}(Q) - \mathcal{T}(Q')\|_\infty \leq \gamma \|Q - Q'\|_\infty.$$

Proof: $\|\mathcal{T}(Q) - \mathcal{T}(Q')\|_\infty$

$$= \gamma \max_{s,a} \left| \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q(s', a') \right] - \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q'(s', a') \right] \right|$$

$$\leq \gamma \max_{s,a} \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left| \max_{a' \in \mathcal{A}} Q(s', a') - \max_{a' \in \mathcal{A}} Q'(s', a') \right|$$

Here, we used the fact $|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)|$.

Contraction of the Bellman's operator

Lemma 5 (γ-contraction of Bellman operator)

For any Q and Q', it holds

$$\|\mathcal{T}(Q) - \mathcal{T}(Q')\|_\infty \leq \gamma \|Q - Q'\|_\infty.$$

Proof:
$$
\begin{aligned}
&\|\mathcal{T}(Q) - \mathcal{T}(Q')\|_\infty \\
&= \gamma \max_{s,a} \left| \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q(s', a') \right] - \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q'(s', a') \right] \right| \\
&\leq \gamma \max_{s,a} \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left| \max_{a' \in \mathcal{A}} Q(s', a') - \max_{a' \in \mathcal{A}} Q'(s', a') \right| \\
&\leq \gamma \max_{s,a} \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} |Q(s', a') - Q'(s', a')|
\end{aligned}
$$

Here, we used the fact $|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)|$.

Contraction of the Bellman's operator

Lemma 5 (γ-contraction of Bellman operator)

For any Q and Q', it holds

$$\|\mathcal{T}(Q) - \mathcal{T}(Q')\|_\infty \le \gamma \|Q - Q'\|_\infty.$$

Proof: $\|\mathcal{T}(Q) - \mathcal{T}(Q')\|_\infty$

$$= \gamma \max_{s,a} \left| \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q(s',a') \right] - \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q'(s',a') \right] \right|$$

$$\le \gamma \max_{s,a} \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left| \max_{a' \in \mathcal{A}} Q(s',a') - \max_{a' \in \mathcal{A}} Q'(s',a') \right|$$

$$\le \gamma \max_{s,a} \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} |Q(s',a') - Q'(s',a')|$$

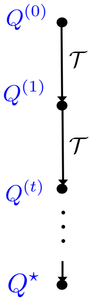
$$\le \gamma \max_{s',a'} |Q(s',a') - Q'(s',a')| = \gamma \|Q - Q'\|_\infty.$$

Here, we used the fact $|\max_a f(a) - \max_a g(a)| \le \max_a |f(a) - g(a)|$.

Value iteration

Value iteration

For $t = 0, 1, \ldots,$
$$Q^{(t+1)} = \mathcal{T}(Q^{(t)})$$



Convergence rate of value iteration

Theorem 6 (Linear convergence of value iteration)

$$\|Q^{(t)} - Q^{\star}\|_{\infty} \leq \gamma^{t} \|Q^{(0)} - Q^{\star}\|_{\infty}$$

- This is implied immediately by the γ-contraction property.

Convergence rate of value iteration

Theorem 6 (Linear convergence of value iteration)

$$\|Q^{(t)} - Q^\star\|_\infty \le \gamma^t \|Q^{(0)} - Q^\star\|_\infty$$

- This is implied immediately by the γ-contraction property.

Implications: to achieve $\|Q^{(t)} - Q^\star\|_\infty \le \epsilon$, it takes no more than

$$\frac{1}{1-\gamma} \log\left(\frac{\|Q^{(0)} - Q^\star\|_\infty}{\epsilon}\right)$$

iterations.

From Q-function to policy

Lemma 7 ([Singh and Yee, 1994])

Let the greedy policy w.r.t. Q *be* π_Q, *then*

$$V^\star - V^{\pi_Q} \leq \frac{2}{1-\gamma} \|Q^\star - Q\|_\infty.$$

From Q-function to policy

Lemma 7 ([Singh and Yee, 1994])

Let the greedy policy w.r.t. Q be π_Q, then

$$V^\star - V^{\pi_Q} \le \frac{2}{1-\gamma}\|Q^\star - Q\|_\infty .$$

$$\boxed{\|\widehat{Q} - Q^\star\|_\infty \le \epsilon} \xrightarrow{\widehat{\pi} = \mathsf{Greedy}(\widehat{Q})} \boxed{V^\star - V^{\widehat{\pi}} \le \frac{\epsilon}{1-\gamma}}$$

- Mind the error amplification factor $\frac{1}{1-\gamma}$

Proof of Lemma 7

Fix state $s \in \mathcal{S}$ and let $a = \pi_Q(s)$. It follows that

$$
\begin{aligned}
&V^\star(s) - V^{\pi_Q}(s) \\
&= Q^\star(s, \pi^\star(s)) - Q^{\pi_Q}(s, \pi_Q(s)) \\
&= \underbrace{Q^\star(s, \pi^\star(s)) - Q(s, \pi_Q(s))}_{=:\mathrm{I}} + \underbrace{Q(s, \pi_Q(s)) - Q^\star(s, \pi_Q(s))}_{=:\mathrm{II}} \\
&\qquad + \underbrace{Q^\star(s, \pi_Q(s)) - Q^{\pi_Q}(s, \pi_Q(s))}_{=:\mathrm{III}}
\end{aligned}
$$

We shall bound each of these terms separately.

- For term I, since $Q(s, \pi_Q(s)) \geq Q(s, \pi^\star(s))$,

$$
\begin{aligned}
Q^\star(s, \pi^\star(s)) - Q(s, \pi_Q(s)) &\leq Q^\star(s, \pi^\star(s)) - Q(s, \pi^\star(s)) \\
&\leq \|Q^\star - Q\|_\infty .
\end{aligned}
$$

Proof of Lemma 7

- For term II,

$$Q(s, \pi_Q(s)) - Q^\star(s, \pi_Q(s)) \leq \|Q^\star - Q\|_\infty.$$

- For term III, by Bellman equations,

$$\begin{aligned} Q^\star(s, \pi_Q(s)) - Q^{\pi_Q}(s, \pi_Q(s)) &= \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\star(s') - V^{\pi_Q}(s')] \\ &\leq \gamma \|V^\star - V^{\pi_Q}\|_\infty \end{aligned}$$

To sum up,

$$\|V^\star - V^{\pi_Q}\|_\infty \leq 2\|Q^\star - Q\|_\infty + \gamma \|V^\star - V^{\pi_Q}\|_\infty$$

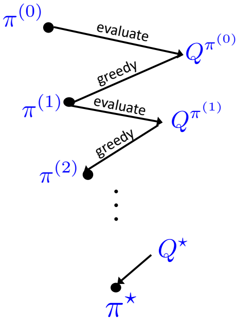
$$\Longrightarrow \qquad \|V^\star - V^{\pi_Q}\|_\infty \leq \frac{2\|Q^\star - Q\|_\infty}{1-\gamma}$$

Policy iteration

Policy iteration

For $t = 0, 1, \ldots,$

$$\pi^{(t)} = \mathsf{Greedy}(Q^{(t-1)})$$
$$Q^{(t)} = Q^{\pi^{(t)}}$$



—*"the dance"*

Convergence rate of policy iteration

Theorem 8 (Linear convergence of policy iteration)

For policy iteration, it follows that

1. $Q^{(t+1)} \geq \mathcal{T}(Q^{(t)}) \geq Q^{(t)}$
2. $\|Q^{(t+1)} - Q^\star\|_\infty \leq \gamma \|Q^{(t)} - Q^\star\|_\infty$

- Policy iteration produces a sequence of improving policies.

Implications: to achieve $\|Q^{(t)} - Q^\star\|_\infty \leq \epsilon$ for output policy $\pi^{(t)}$, it takes no more than

$$\frac{1}{1-\gamma} \log \left(\frac{\|Q^{(0)} - Q^\star\|_\infty}{\epsilon} \right)$$

iterations.

Proof for policy iteration: policy improvements

Proof of $\mathcal{T}(Q^{(t)}) \geq Q^{(t)}$:

$$\mathcal{T}(Q^{(t)})(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^{\pi^{(t)}}(s',a')$$

Proof for policy iteration: policy improvements

Proof of $\mathcal{T}(Q^{(t)}) \geq Q^{(t)}$:

$$\begin{aligned}
\mathcal{T}(Q^{(t)})(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^{\pi^{(t)}}(s',a') \\
&\geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} Q^{\pi^{(t)}}(s', \pi^{(t)}(s'))
\end{aligned}$$

Proof for policy iteration: policy improvements

Proof of $\mathcal{T}(Q^{(t)}) \geq Q^{(t)}$:

$$\begin{aligned}
\mathcal{T}(Q^{(t)})(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^{\pi^{(t)}}(s',a') \\
&\geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} Q^{\pi^{(t)}}(s', \pi^{(t)}(s')) \\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^{\pi^{(t)}}(s') = Q^{\pi^{(t)}}(s,a).
\end{aligned}$$

Proof for policy iteration: policy improvements

Proof of $\mathcal{T}(Q^{(t)}) \geq Q^{(t)}$:

$$\begin{aligned}
\mathcal{T}(Q^{(t)})(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^{\pi^{(t)}}(s', a') \\
&\geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} Q^{\pi^{(t)}}(s', \pi^{(t)}(s')) \\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^{\pi^{(t)}}(s') = Q^{\pi^{(t)}}(s,a).
\end{aligned}$$

Proof of $Q^{(t+1)} \geq \mathcal{T}(Q^{(t)})$: From policy improvement theorem, we already know $Q^{(t+1)} \geq Q^{(t)}$.

Proof for policy iteration: policy improvements

Proof of $\mathcal{T}(Q^{(t)}) \geq Q^{(t)}$:

$$
\begin{aligned}
\mathcal{T}(Q^{(t)})(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^{\pi^{(t)}}(s',a') \\
&\geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} Q^{\pi^{(t)}}(s', \pi^{(t)}(s')) \\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^{\pi^{(t)}}(s') = Q^{\pi^{(t)}}(s,a).
\end{aligned}
$$

Proof of $Q^{(t+1)} \geq \mathcal{T}(Q^{(t)})$: From policy improvement theorem, we already know $Q^{(t+1)} \geq Q^{(t)}$.

$$
\begin{aligned}
Q^{\pi^{(t+1)}}(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} Q^{\pi^{(t+1)}}(s', \pi^{(t+1)}(s')) \\
&\geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} Q^{\pi^{(t)}}(s', \pi^{(t+1)}(s'))
\end{aligned}
$$

Proof for policy iteration: policy improvements

Proof of $\mathcal{T}(Q^{(t)}) \geq Q^{(t)}$:

$$
\begin{aligned}
\mathcal{T}(Q^{(t)})(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^{\pi^{(t)}}(s',a') \\
&\geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} Q^{\pi^{(t)}}(s', \pi^{(t)}(s')) \\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^{\pi^{(t)}}(s') = Q^{\pi^{(t)}}(s,a).
\end{aligned}
$$

Proof of $Q^{(t+1)} \geq \mathcal{T}(Q^{(t)})$: From policy improvement theorem, we already know $Q^{(t+1)} \geq Q^{(t)}$.

$$
\begin{aligned}
Q^{\pi^{(t+1)}}(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} Q^{\pi^{(t+1)}}(s', \pi^{(t+1)}(s')) \\
&\geq r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} Q^{\pi^{(t)}}(s', \pi^{(t+1)}(s')) \\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^{\pi^{(t)}}(s',a') = \mathcal{T}(Q^{(t)}).
\end{aligned}
$$

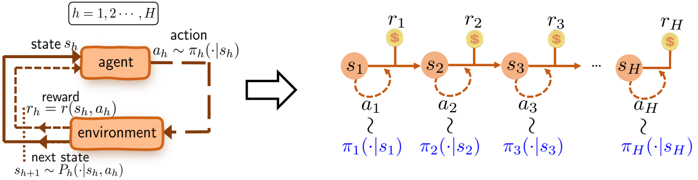
Proof for policy iteration: linear convergence

Using $Q^{(t+1)} \geq \mathcal{T}(Q^{(t)})$,

$$\begin{aligned} \|Q^\star - Q^{(t+1)}\|_\infty &\leq \|Q^\star - \mathcal{T}(Q^{(t)})\|_\infty \\ &= \|\mathcal{T}(Q^\star) - \mathcal{T}(Q^{(t)})\|_\infty \\ &\leq \gamma \|Q^\star - Q^{(t)}\|_\infty . \end{aligned}$$

Here, the last line follows from the contraction of the Bellman's optimality operator.

Bellman's optimality eq. for finite-horizon MDPs



Let $Q_h^\star(s,a) = \max_\pi Q_h^\pi(s,a)$ and $V_h^\star(s) = \max_\pi V_h^\pi(s)$.

1. Begin with the terminal step $h = H+1$:

$$V_{H+1}^\star = 0, \quad Q_{H+1}^\star = 0.$$

2. Backtrack $h = H, H-1, \ldots, 1$:

$$Q_h^\star(s,a) := \underbrace{\mathbb{E}\left[r_h(s_h,a_h)\right]}_{\text{immediate reward}} + \underbrace{\mathbb{E}_{s' \sim P_h(\cdot|s,a)} V_{h+1}^\star(s')}_{\text{next step's value}}$$

$$V_h^\star(s) := \max_{a \in \mathcal{A}} Q_h^\star(s,a), \qquad \pi_h^\star(s) = \text{argmax}_{a \in \mathcal{A}} Q_h^\star(s,a).$$

References I

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. (2019).
Reinforcement learning: Theory and algorithms.

Bellman, R. (1952).
On the theory of dynamic programming.
Proceedings of the National Academy of Sciences of the United States of America, 38(8):716.

Singh, S. P. and Yee, R. C. (1994).
An upper bound on the loss from approximate optimal-value functions.
Machine Learning, 16:227–233.